Dynamic Neural Network-based Solutions for Acoustic Echo Suppression

Jin Li [a], Sijie Liu [b], Yafeng Wu [c, *]

School of Power and Energy, Northwestern Polytechnical University, Xi'an 710072, China

[a] jmin120303@foxmail.com, [b] sijieliu_123@sina.com, [c, *]yfwu@nwpu.edu.cn

* Corresponding Author Email: yfwu@nwpu.edu.cn

Abstract. Adaptive filters are widely used as core algorithms in current acoustic echo cancellation (AEC) systems. Echo path estimation is carried out by using different adaptive strategies in the adaptive filter. When nonlinear echo occurs, the performance deteriorates and seriously affects the call quality of both terminals. Taking advantage of the excellent non-linear fitting ability of neural networks in this paper, dynamic neural networks with self-updating parameters work continuously during the inference stage. The algorithm was computed and evaluated using publicly available echo audio data, showing that the dynamic neural network performs approximately as well as the optimal algorithm in the linear echo environment, and outperforms existing algorithms in the non-linear echo environment.

Keywords: Acoustic Echo Suppression; Non-linear Echo; Dynamic Neural Networks.

1. Introduction

The phenomenon of acoustic coupling is widespread in instant messaging systems consisting of loudspeaker microphones and has been a continuing concern for decades[1,2,3]. Impact communication is a typical manifestation of the consequences for acoustic echoes. Initially perceived by the general public in the telephone network, there are two types of echoes in this scenario: the former is the hybrid echo caused by an unbalanced hybrid bridge in the circuit and the latter is the acoustic echo caused by the microphone receiving sound directly or indirectly from the speaker.

The adaptive filter algorithm is one of the most widely used in the echo cancellation problem, with a large variety that is rapidly becoming mainstream after an extreme historical phase. However, the non-linear, time-varying and unstable nature of the echo path is exacerbated by equipment quality issues and more interference with the communication channel [4]. The algorithm faces a great challenge, and research into high-performance acoustic echo suppression algorithms remains a valuable area of research in the field of speech signal processing. Neural networks go through three stages before taking into practice: the network model building and parameter initialization the parameter training and model inference, and the deployment. It can be divided into static and dynamic networks in terms of parameter and structural changes. The structure of the network and some of the parameters change during the inference process stage is characteristic of dynamic networks[5,6]. Compared to static networks, dynamic networks enhance the adaptability to different data and balance the tension between computational accuracy and speed. Breakthroughs in network inference efficiency and expressiveness dynamism are the focus of dynamic neural network research, and structural adaption and parameter adaptive studies have been focused on by many researchers[7]. Adaptive filter theory and dynamic neural networks are considered in this paper. An extremely effective structure has been designed with excellent results.

We first organized the structure of the process using the method, then designed a dynamic neural network update structure, and finally combined the complete neural network. The training process started with parameter pre-training, and details were adjusted during inference, resulting in superior results to existing methods.

1.1 Related Work

In recent years, the non-smoothness and non-linear characteristics of the echo path have attracted the attention of researchers, giving rise to numerous non-linear residual echo suppression algorithms [8,9]. These algorithms sacrifice algorithmic complexity to obtain limited performance gains, which are based on improvements to the adaptive filters. Adaptive filters are, of course, still the dominant algorithm for this problem, and a two-step approach is commonly used: first a frequency-domain adaptive filter is used to remove linear echoes; then a post-filter is applied in the subsequent processing stage for residual echo suppression(RES)[10]. The literature on echo suppression using neural networks has grown in the last decade[11]. Although good results have been achieved, however, changing echo paths is ignored in such studies and echo paths are constantly changing in real environments is defied. Therefore, another aspect, which uses neural networks to suppress residual echoes and replace the post filter, has achieved some results[12,13], evading the shortcomings of neural networks. This problem is well solved because the dynamic neural network concept, pre-exploratory work in previous studies, has produced many well-constructed structures.

In summary, the evidence presented in this section suggests that: the echo suppression problem is a strongly non-linear data mapping problem; the parameters of the algorithm need to have adaptive updating capabilities because of the time-varying; and dynamic neural networks have great potential for solving this problem.

1.2 Mathodology

The algorithm flow diagram is shown in Figure 2: the near-end audio and the far-end audio are processed on frame and window, with the near-end audio entering the delay estimation module and the far-end audio entering the history frame, both of which are stored as corresponding parameter training data. When the error exceeds the limit, the training module is activated, the minimum mean square error is calculated and the parameters are updated, and then the parameters are updated to the dynamic neural network to complete the real-time estimation of the echo path. The specific parameter settings are described in detail below.

The Elman neural network and the NARX neural network are referencing in this paper because of the feedback characteristics of acoustic echoes that are considered. The use of wavelet neural networks for the time-frequency domain characteristics of the audio signal. This is replaced by a wavelet basis function by performing a network topology transformation at the nodes out.

The wavelet basis function used in this paper is the Morlet mother wavelet basis function, the expression of which is:

$$y = \cos(1.75x)e^{-x^2/2} \tag{1}$$

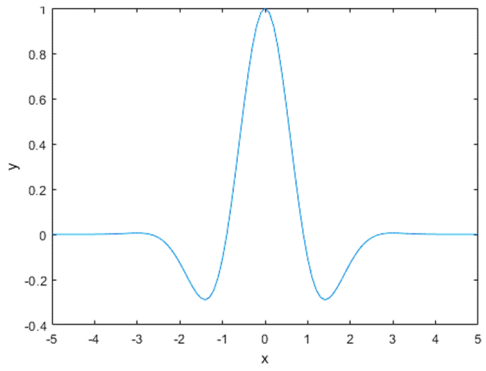

Fig 1. This is the wavelet basis function image

As shown in Figure 3, the structure diagram of the dynamic neural network used in this paper. The network consists of a dynamic part and a static part, where the static network guarantees the basic computation of the trained environment and the dynamic part guarantees the effective tracking of the

real environment changes. Pre-training is introduced to determine the parameters of network layers 1 and 2 , and the last two layers are updated using the mean square error as the loss function, so that the parameters are continuously updated in the inference process in the reverse direction.

The speech signal and echo signal use a frame-splitting strategy, entering the calculation module by frame, and first performing the near-end signal and far-end signal delay estimation. A finite length of far-end signal data is cached, and this data retention length needs to include the maximum time when the echo occurs in the system. The signal stream is saved by frame as the original corpus. The signals obtained after delay estimation are saved against each other as the target signals, and a separate section of memory is divided and stored according to a stack, where new signals are continuously entered and saved, and old signals are continuously deleted. In the same way as the adaptive filter algorithm, the parameters are fixed until the computational error does not exceed a limited value. If the algorithm determines that the parameters need to be updated, it will perform a feedback update that is shared with the computational network parameters in real time, which is essentially the same as the adaptive filter that tracks signal changes for parameter updates.

By combining the network infrastructures, it can meet the conditions of small number of parameters, fast convergence, and low complexity applications while avoiding the difficulties of real-time computation caused by oversized network structures. At the same time the computational process with nonlinear adaptability. Using mean squared error as a dynamic update criterion, with minimum mean squared error as the ultimate goal. As a commonly used loss function, the formula is susceptible to the use of gradient descent algorithms. However, this algorithm is too sensitive to outliers, causing the final model effect to be biased due to outliers. Alternatives are to use mean absolute error, L1 parametric, L2 parametric loss functions, etc.

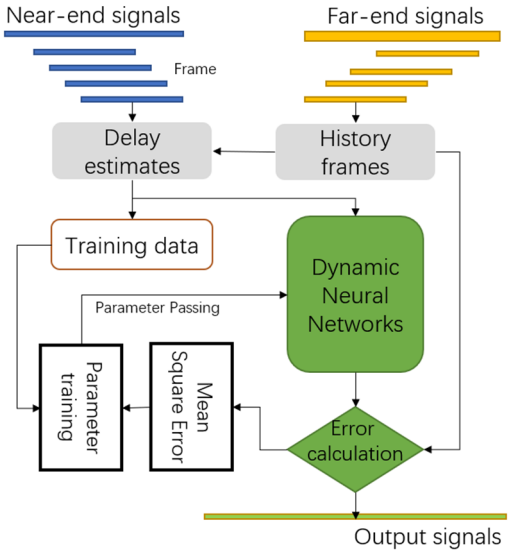

Fig 2. Algorithm flow chart

By profiling existing networks with dynamic parameter capabilities, the main structures of the three neural networks can be extracted. By deepening the number of layers and changing the calculation method, the advantages of each algorithm can be adopted to achieve the final goal.

The process of calculating the hidden and take-up layers can be expressed as follows:

$$x(k) = f(w_1 x(k-1)) + w_2(u(k-1)) \tag{2}$$

where x is the node unit vector, $f(*)$ is the transfer function, w_1 is the connection weight between the takeover layer and the hidden layer, and w_2 is the connection weight between the hidden layer and the input.

Finally, the output signal is plugged into the hidden layer, i.e. the feedback is plugged into the input during signal transmission, and a time delay and output feedback mechanism is added to the neural network. This method is widely used in non-linear sequence prediction problems.

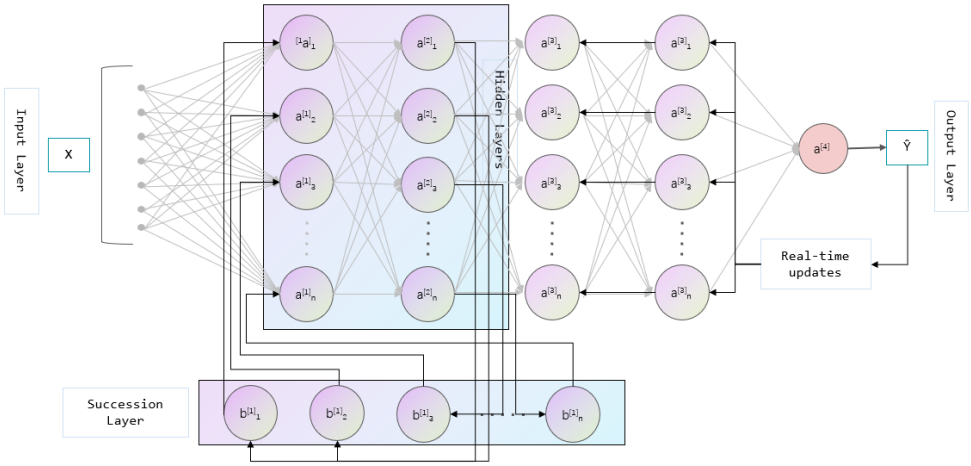

Fig 3. Dynamic neural network

2. Experimental Evaluation

After the above description, a basic description of the problem context and algorithm is given. Next an evaluation test of the algorithm is carried out, the dataset and metrics are presented and the effectiveness of the algorithm is demonstrated through a comparison of multiple algorithms.

2.1 Dataset

We tested the algorithm on the ICASSP 2021 dataset[14], a collection of 2500 different real environments, audio devices and human speakers. The data was obtained through large-scale crowdsourcing. Each audio set contains a distal single-speaker situation, a proximal single-speaker situation and a dual-speaker situation, where the echo path is changed by instructing the user to move around the device or to allow themselves to move the device. A time domain graphical representation of one of the audio sets is shown in Figure 4.

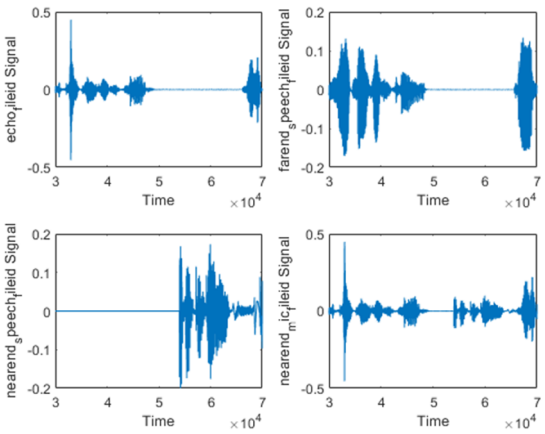

Fig 4. Example of an audio time domain diagram (*top left is the echo signal; *top right is the far-end voice signal;* bottom left is the near-end voice signal; *ranger is the near-end mixed signal)

2.2 System Parameter Setting

The sampling rate of the signals used in the experiments was 16Khz, the input signal of the system was framed with a frame length of 20ms and a frame shift of 10ms, and the window function used is

the Blackman window. After that, a wavelet transform is applied to each frame, which gives a time-frequency representation of the input signal, where each frame has 161 frequency points. The dimensionality of the input features is therefore 161*2=322.

In the network, the Relu function is used for the activation function, the wavelet basis function described by Equation x is used for the node calculation, and the Adam optimizer is used for training in the pre-training phase, with MSE as the loss function and a learning rate of 0.0003, trained 200 times on all training samples.

2.3 Results of the Experiment

The evaluation metrics for the tests were Echo return loss enhancement (ERLE, Equation 3), Perceptual evaluation of speech quality (PESQ) and signal-to-distortion ratio (SDR, Equation 4).

$$ERLE = 10\lg\frac{E[d^2(n)]}{E[e^2(n)]} \tag{3}$$

$$SDR = 10\lg\left(\frac{P_{Signal}}{P_{Noise} + P_{Distortion}}\right) \tag{4}$$

The experimental results are shown in Table I and Table II, comparing the Wiener filter, the frequency domain chunked adaptive filter, the multilayer CNN neural network and the dynamic neural network. From the conclusions, it shows that the dynamic neural networks all achieve more stable results in both linear and non-linear echoes, with a slight improvement over the frequency domain chunking adaptive filters which are widely used in the industry. The Wiener filter is relatively insignificant as a primary adaptive filter method. the CNN network has a significant improvement for targeted training audio, but performs poorly and has poor robustness for other audio.

Two different comparison scales were set, one for time and the other for signal-to-noise ratio when conducting the algorithm experiments. It can be seen that on the time scale, the longer the time period will give a small general improvement in the algorithm's effectiveness. On the signal-to-noise scale, the lower the signal-to-noise ratio, the less effective the algorithm is at processing. When compared under the same conditions, each algorithm shows effectiveness in a linear echo environment, with the dynamic neural network algorithm performing similarly to the frequency domain chunking adaptive filter algorithm.

Table 1. Average results of different algorithms in a linear echo environment

		ERLE		**PESQ**		**SDR**	
		10s	*30s*	*10s*	*30s*	*10s*	*30s*
3.5dB	**WF**	16.11	19.21	1.12	1.25	5.61	5.34
	CNN	30.25	32.24	1.26	1.34	5.39	5.67
	FDBA	46.25	44.31	1.54	1.36	**6.59**	6.24
	DYNN	**49.56**	**48.53**	**1.59**	**1.68**	6.58	**6.94**
7dB	**WF**	17.32	18.95	1.24	1.33	5.67	5.14
	CNN	20.57	21.17	1.21	1.34	6.11	5.12
	FDBA	45.21	55.36	**1.56**	**1.55**	**6.56**	6.34
	DYNN	**49.91**	**56.29**	1.35	1.52	6.21	**6.54**

(*Note: "WF" refers to Wiener filter, "CNN" refers to Convolutional Neural Network, "FDBA" refers to Frequency Domain Blocking Adaptive Filter, "DYNN" refers to is the dynamic neural network used in this paper.)

The pattern that emerges is consistent, with non-linear and linear echo environments, in time and signal-to-noise ratio. In particular, the dynamic neural network metrics were much better than the other algorithms. To be precise, performance degradation occurs in all algorithms when dealing with non-linear echoes. The dynamic neural network has a much smaller performance degradation and can therefore cope effectively with non-linear echoes.

Table 2. Average results of different algorithms in a No-linear echo environment

		ERLE		**PESQ**		**SDR**	
		10s	*30s*	*10s*	*30s*	*10s*	*30s*
3.5dB	**WF**	11.35	11.28	0.64	0.59	5.12	4.99
	CNN	19.35	20.64	1.02	1.09	5.99	5.87
	FDBA	26.34	21.51	1.31	1.10	6.01	6.23
	DYNN	**40.21**	**42.61**	**1.49**	**1.59**	**6.99**	**6.57**
7dB	**WF**	12.35	11.54	0.81	0.69	5.10	5.12
	CNN	22.21	26.31	1.21	1.19	6.21	6.15
	FDBA	23.34	23.54	1.20	1.19	5.95	5.69
	DYNN	**44.59**	**43.21**	**1.54**	**1.64**	**6.15**	**6.66**

(*Note: "WF" refers to Wiener filter, "CNN" refers to Convolutional Neural Network, "FDBA" refers to Frequency Domain Blocking Adaptive Filter, "DYNN" refers to is the dynamic neural network used in this paper.)

The pattern that emerges is consistent, with non-linear and linear echo environments, in time and signal-to-noise ratio. In particular, the dynamic neural network metrics were much better than the other algorithms. To be precise, performance degradation occurs in all algorithms when dealing with non-linear echoes. Dynamic neural networks have a much smaller performance degradation and can therefore cope effectively with non-linear echoes. The spectrograms in Figure 5, show different algorithms processing the same set of audios.

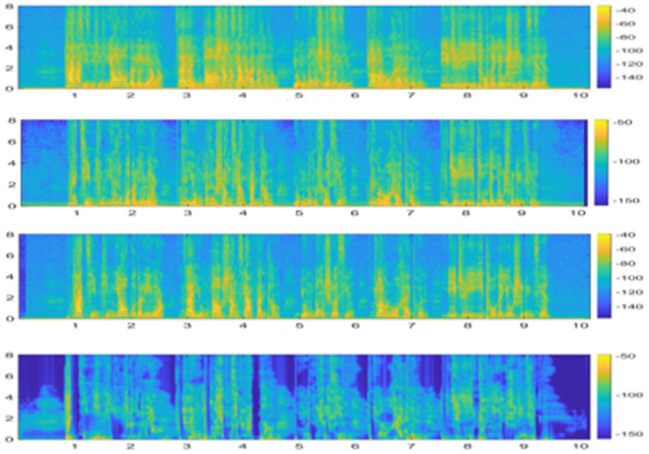

Fig 5. Sound spectrum after algorithm processing(The horizontal coordinate is time(s); the vertical coordinate is frequency(Hz).)

3. Conclusion and Next Steps

In this study, a dynamic parametric neural network algorithm is proposed in order to solve the non-linear acoustic echo problem. A dynamic parametric network layer that can track the echo path is

designed to cope with the time-varying characteristics of the echo path, making full use of the powerful non-linear adaptation capability of the neural network. Combining Elman neural network, wavelet neural network and NARX neural network, its core computational nodes are stripped from its network theory analysis. The network is formed by pairing the core computational nodes. The algorithm was evaluated with real echo data, and it was concluded that the algorithm can obtain similar results to the frequency domain chunking adaptive filtering algorithm widely used in industry when dealing with linear data, and achieved even better simulation results on non-linear data, fully demonstrating the effectiveness of the algorithm.

Further future work could be carried out in the following areas: 1) Finding more clever methods of updating dynamic parameters to further improve the tracking of the algorithm. 2) In the course of the work carried out, we found that there is high scope for improving the performance of the echo detector, which is one of the reasons why the experimental performance is not as good as the simulation performance. 3) More extensive environmental experiments are needed to further test the performance of the algorithm.

References

[1] Gustafsson S, Martin R, Jax P, et al. “A psychoacoustic approach to combined acoustic echo cancellation and noise reduction”. IEEE Transactions on Speech and Audio Processing, 2002, 10(5): 245-256.

[2] Reuven G, Gannot S, Cohen I. “Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller”. Speech communication, 2007, 49(7-8): 623-635.

[3] Schwarz A, Hofmann C, Kellermann W. “Combined nonlinear echo cancellation and residual echo suppression”. Speech Communication; 11. ITG Symposium. VDE, 2014: 1-4.

[4] Gao Y, Liu I, Zheng J, et al. “Independent Echo Path Modeling for Stereophonic Acoustic Echo Cancellation”. INTERSPEECH. 2020: 3955-3958.

[5] Wu D, Pigou L, Kindermans P J, et al. “Deep dynamic neural networks for multimodal gesture segmentation and recognition”. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(8): 1583-1597.

[6] Han Y, Huang G, Song S, et al. “Dynamic neural networks: A survey”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[7] Xu Z, Li S, Zhou X, et al. “Dynamic neural networks based kinematic control for redundant manipulators with model uncertainties”. Neurocomputing, 2019, 329: 255-266.

[8] Pfeifenberger L, Pernkopf F. “Nonlinear Residual Echo Suppression Using a Recurrent Neural Network”. Interspeech. 2020: 3950-3954.

[9] Cohen A, Barnov A, Markovich-Golan S, et al. “Joint beamforming and echo cancellation combining QRD based multichannel AEC and MVDR for reducing noise and non-linear echo”. 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018: 6-10.

[10] Ivry A, Cohen I, Berdugo B. “Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression”. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 126-130.

[11] Zhang H, Wang D. “Deep learning for acoustic echo cancellation in noisy and double-talk scenarios”. Training, 2018, 161(2): 322.

[12] Wang Z, Na Y, Liu Z, et al. “Weighted recursive least square filter and neural network based residual echo suppression for the aec-challenge”. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 141-145.

[13] Zhang S, Wang Z, Sun J, et al. “Multi-task deep residual echo suppression with echo-aware loss”. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 9127-9131.

[14] Sridhar K, Cutler R, Saabas A, et al. “ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results”. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 151-155.